Deloitte.

The Deloitte On Cloud Podcast

David Linthicum, Managing Director, Chief Cloud Strategy Officer, Deloitte Consulting LLP

Title: **How companies are capitalizing on opportunities to increase their cloud value**

Description: Cloud brings tremendous innovation opportunities. In this episode, VMware's William Lam talks with David Linthicum about the latest trends he's is seeing in the cloud space—such as the rise in hybrid and multi-cloud, the modernization of on-prem architectures, disaster recovery, and the explosion of data and AI for insight-driven decision-making. The two also discuss how companies can leverage these trends at scale with sound planning and by optimizing cloud ecosystem connectivity.

Duration: **00:24:16**

David Linthicum:
Welcome back to the On Cloud podcast. Today on the show I'm joined by William Lam, senior staff solution architect with VMware. Welcome to the show, William.

William Lam:
Thanks for having me, David.

David Linthicum:
Yeah. It's great to have you. I was just on your podcast that you guys do, so I appreciate that. I'm just really looking to catch up on where this market is going. Give us the William Lam story. Where did you come from? What were you doing before you got to VMware? What do you focus on in your career?

William Lam:
Thanks again for having me. My name is William Lam. Prior to VMware, I was a customer—you know, I've been fortunate to work for a lot of large enterprises. Prior to joining VMware, I was at Salesforce, on the infrastructure structure there, which of course is made up of VMware. I've been using our software stack roughly since around the 2008 timeframe.

As I mentioned, I'm a senior staff solutions architect, part of our VMware cloud infrastructure business group. I focus on working with our customers, product management, and continue to focus on our hybrid and multi-cloud solutions. I'm coming up actually on 12 years this coming fall. So, it feels like a lifetime. I get to work day-to-day with our customers or engineering teams, and I really try to distill down the requirements and work collaboratively across our different organizations, kind of deliver on this vision of the VMware cloud to our customers, whether that's private cloud, some of our partner clouds, even all the way out on the edge.

David Linthicum:
What's the most enjoyable aspect of your job?

William Lam:
Really, just talking to our customers and hearing about the different use cases. I've been very fortunate to be able to touch a lot of our platforms, from the core hypervisor, storage, networking, and the management of products. There is not a time that goes by, where I don't get to talk to our large and strategic customers and really hear about how they're pushing out our software stack or services to the boundaries, and doing some really, really interesting things.

You can only imagine some of the stories I hear about where are they trying to run our software stack, especially at the edge. I did a blog post not too long ago where a customer shared, "They've actually got it deployed in a Ferrari." I was just like, "Wow." I was just asking, "Hey, where was the craziest place you've run the hypervisor?" And I had all these customers share their stories. It really is true that we have our software stack in all the different vertical segments. You name it, ships, boats, submarines, whatever it may be. It's just really humbling to hear about how our customers are able to leverage our technology to really transform their organization and deliver their business value to their end customer. So, that has always been something that I enjoy doing, speaking with our customers.

David Linthicum:
Yeah. Edge computing makes things more fun because we're not just staring at, or dealing with, centralized computing. We're dealing with different devices and things like that. I've always enjoyed that aspect of computing. So, let's talk about hybrid and multi-cloud, complex cloud architectures. Let's get the definitions out there first. Tell the audience, and I know most people who listen to the podcast are going to know what a hybrid cloud and a multi-cloud is, but how are you guys defining it at VMware?

William Lam:
It's interesting because a couple years back, when the term multi-cloud came out, VMware was starting to talk a little bit about it. A lot of this was early insights from some of our strategic customers, some of the challenges that they were faced with. It was sort of seen as, "Hey, this is just the next new buzzword from cloud, and we just needed something to talk about." If you fast-forward to today, I think in some of our reports where we've done surveys with third parties, and I think even HashiCorp published the state of multi-cloud for the 2023 revision.

I think it's like 83 percent of organizations today have multi-cloud, anywhere from one to two or more than that. We see a lot of that certainly driven around the time of the pandemic. IT organizations were going through a transformation, and they were trying to deliver certain capabilities and reacting to what was happening around the world, and really figuring out where best to put these workloads. So, we started to see that come up for a lot of the larger customers, and deploying the workload there based on the needs and requirements. At the same time, we have a huge on-premise customer user base, and that's where VMware started from. Over the last several years, we clearly see that it's not just on-prem versus public cloud versus partner cloud. It really is turning into this distributed hybrid infrastructure that our customers are tasked with managing, operating at scale. These are the sorts of challenges that we as an industry are faced with. And how is VMware helping their customers with the different technologies and solutions that we have to offer?

David Linthicum:
Yeah. I think the common pattern is we're moving to complex cloud deployments. That's typically going to be multiple public clouds, but also even include the existing legacy systems, private clouds you may have around, and even edge-based systems. So, when you look at it, we're kind of solving the same problem. You have processes in different heterogenous cloud environments, and different platforms we're working with on-premise. That's really the architecture of where we are now. When you think of moving to multi-cloud, people got to multi-cloud—and you alluded to this in your response—this was something where it was really chasing best-of-breed technology.

It occurred organically in the past, in other words, "I want to use this particular AI platform and this particular database, and it happens not to be on this particular public cloud provider, and, by the way, I need to use my existing legacy stuff. I need to use edge-based stuff." So, we're getting to a point where we're going to be dealing with huge amounts of heterogeneity, different cloud providers that we're dealing with. Really, this is something that's moving up from the cloud provider into this kind of abstract space. People are calling it supercloud and metacloud. I'm just saying focusing on cross-cloud services ultimately is a way to mitigate the risk and increase the productivity of leveraging hybrid and multi-clouds. What are your thoughts?

William Lam:
I completely agree. It definitely happened organically. What we see from our customers is that the organizations that were really well hooked into their business operations really understood that you needed to be differentiated. The pandemic forced overall adoption of public cloud. You went from an organization that was 100 percent in the office to now I've got to deliver IT services remotely globally. How do I do that? The only way I can do that with any kind of scale was leveraging public cloud, and we certainly saw an influx of our customers overnight go from purely 100 percent on-premises to now delivering cross-cloud services across their different regions. How do you start to manage that?

So, we saw this huge growth of consumption of all these different services. Now, what we're seeing is that with the macroeconomics, with the focus of really turning IT from a cost center and really turning it into an enabler, we do have net new challenges I can clearly outline. There are now new constraints and requirements on where work was run. One big one that we see a lot of is around data sovereignty and compliance. This has been there for quite some time, and this is where VMware has done some work with our sovereign cloud partners. Last year, we announced that at VMware Explore. We're seeing that it's not just best-of-breed clouds, but we're also seeing partner clouds being able to meet certain types of requirements.

So, our customers are now saying, "Not only do I need to run these workloads in a particular region, but I also need the folks that are staffing that to be local versus being remote." So, there's all these new requirements that come in that just create additional challenges. How do you manage this at scale, especially from an operations standpoint? So, cost becomes a big factor from the organization to take a look at their overall deployment, whether it's multi-cloud or on-prem, and saying, "Where best should I start to think about these workloads? How do I start to move them around, or strategically place workloads where they need to be at with this additional constraints in place?"

David Linthicum:
So, you're talking to a CIO and they're moving into a hybrid or multi-cloud environment, a complex cloud environment. What are some of the challenges in terms of adoption? What should they be focused on? Where is this market actually moving to from where we are now in 2023 to where we're going to be in 2024 and 2025?

William Lam:
There are a couple of key use cases that we continue to still hear, and these are the bread-and-butter use cases that we heard in the very early days of multi-cloud. They just continue on, like this whole concept of IT transformation, digital transformation, nothing new I'm sure to your audience. We just continue to evolve. I'm sure everybody is tired of hearing about AI, but it really is that next thing that's really driving further digital transformation. So, from these executives, they have to think about their overall product portfolio, how they're trying to differentiate.

So, we see a lot of customers still in the typical, "Hey, I'm going to think about modernization across the board." Typically when we talk about modernization, we talk about cloud native, microservices, all these fancy terms about modernizing an application, but one of the things that we don't really talk a lot about, or at least I don't always hear others talk about, is: What about modernizing your infrastructure? We have a lot of customers that run typical on-prem infrastructure. This is aging hardware. Now it's easy to use the word legacy, but if you look at a lot of the processes and what actually runs our overall economy, it's still running on most of these traditional-based applications.

So, while it is "legacy," this is really where a lot of the business value is at. And how can organizations start to modernize their infrastructure, so that they don't actually have to manage this hardware, which can be several years old, especially when we had the supply chain challenges a couple years back? We're only just getting out of that state. So, getting access to modern hardware was extremely difficult. We have a number of customers that feel like running a datacenter is not their bread-and-butter. They want to be able to provide different capabilities, and to be able to do that you need to have modern infrastructure that can provide developer services, run new applications, things like AI, databases and servers, et cetera.

So, they're actually moving their infrastructure into the public cloud or into a partner cloud. We see that in several forms. We see the severe or extreme cases, which is, "I'm completely exiting the datacenter." So, these are our typical datacenter evacuations, and we have a number of large customers already doing that today and running very successfully. Then we have in our mix customers who are doing consolidation. M&A was certainly huge at one point, especially in the IT industry. So, we see a lot of contraction and even expansion, where we've seen some acquisitions just not go through, and they were having to basically divest the organization and say, "How do I start to separate out these different organizations?" So, various forms of datacenter contraction as well as expansion or even consolidation, we see that happening across these different clouds. So, again, kind of this hybrid model.

Then the other one that we see quite a bit is around disaster recovery. With this age of malware and ransomware attacks, it's critical for organizations to really have a true DR scenario, not just for typical, "Hey, my datacenter is no longer available." But when these attacks happen and, "Do I have a good DR strategy?" So, ransomware prevention and recovery is a key one. Then certainly the ability to extend functionality from your datacenter, especially when you look at folks experimenting in the AI/ML space. It's difficult to get hardware such as GPU. So, being able to run an experiment, go out to a public cloud provider, test that scenario, build out that model, and then bring that back into the on-prem datacenter is another popular scenario that we see with a lot of our customers.

David Linthicum:
Yeah. I think what you're getting to is something I've been writing and speaking about a lot. We're looking to put it on the most optimal platform. Whether that's cloud or on-premise or an edge-based system, even if we have to move things around to get to a more optimal state, that's really the target here. So, it's never going to be extreme. It's never going to be all cloud. It's never going to be all datacenter. It's never going to be all edge computing. Will probably never be. Ultimately, we need to back up a bit and understand what those applications are doing, what their performance patterns are, and, ultimately, where they need to run best, and understand we're going to make a few mistakes.

But that really should be the question. What I saw during the pandemic is that people moved very quickly into the cloud, and in many instances didn't do the planning and assessment needed to understand the data and applications that are existing on particular platforms, including the cloud. Now we're kind of moving things around, wouldn't you think?

William Lam:
Yes. Absolutely. I think that's one of the things that we see between a customer who is successful in adopting multi-cloud and being able to take advantage of all these cloud services. That's one of the biggest benefits. You're now able to take these workloads, which typically were behind the corporate firewall, and you're able to bring it closer to these public clouds. The work that we've done brings these workloads closer to these public cloud services.

But one of the things that we found between a customer that was successful versus ones that were challenged was how much planning did they do. Now we understand that there are some customers that are, "Hey, I've got to get out of my datacenter lease. It's a co-lo." We do see those customers and we're able to get them out as quickly as possible. But once you think beyond that, it's like, "What do I do with these workloads?" I'm sure CIOs and CTOs are thinking about, "How do I become differentiated?" The first piece is really just that analysis of your workloads, and we have a lot of tools that we can help you understand that traffic. Networking is a big topic. In a typical datacenter, traffic is typically busy between east and west, and your north/south traffic isn't really exerted that much.

But when you go to the public cloud and you start to think about hybrid, all of a sudden your edge endpoints are starting to get stressed to the point that you might not have been planning on. You might have workloads that get moved to a public cloud, but you've got this database that might be stranded on-prem because of compliance purposes, and all of a sudden you're tromboning traffic. So, we do see customers who don't do that planning upfront have further challenges with it. Then furthermore, we typically talk to these extremes, and maybe it makes a great piece for the media to talk about, but I can still recall these extremes in the past couple of examples.

We talked about bare metal versus hypervisor. There was one point that was like you can never get the performance running as a hypervisor or the VM. We moved away from that and we do see that there's value in scenarios where you might have to have bare metal. But we do believe that most customers, more than 90-plus percent, virtualize. Then we got to the days of VM versus container, like, "Hey, everything is going to be a container." Then we quickly pulled back and we said, "You know what? It doesn't make sense." In some cases it absolutely does not make sense when we look data services, and now we have stateful and stateless applications, and we understand that there are more security concepts built into a virtual machine construct. There's benefits to containers. We harmoniously now understand as an industry that we do see both workloads being deployed evenly, and we want to put the best type of application architecture based on the format.

Then we kind of did this again with private versus public cloud. All work was going to go to a public cloud. The on-prem datacenter was completely dead. I think we do push it to these extremes. Ultimately, VMware was really pragmatic at this approach. It was really listening to our customers and figuring out where the trend is heading. We really do feel that where we're at right now in is that we really are, our customers, living in this hybrid distributed cloud, and our role in this place is really to provide those tools and services, to provide that unified interface, so that our customers can get the benefits of the cloud no matter where their work was running.

David Linthicum:
I love the fact that you take a pragmatic view of that. In other words, it's never going to be that if you have a hammer and everything doesn't look like a nail. At the end of the day, you have to understand what that nail is and get it implemented. Also, I love the term tromboning. I've never heard that before. Is that when curves go up and down, when traffic goes up and down?

William Lam:
Yeah, it just goes up and down. I talk to a lot of our networking colleagues and that's something that they're really concerned about, because all of a sudden you're starting to exercise parts of your infrastructure that typically weren't exercised when you only had to deal with an on-prem datacenter. But once you start to bridge that gap and you want to have true hybridity, you want to have workload mobility, you want to have connectivity, networking is such a fundamental piece of it. It's not something that is always foremost. You don't think about it because it's just plumbing. How hard can it be to connect Point A to Point B?

But it's extremely difficult, and there are so many different technologies, not only within the datacenter that you need to intersect with, but there's also different types of connectivity that the public cloud providers provide. How do you provide access, high bandwidth/low latency? So, there's really some interesting work of the hyperscalers along with VMware over the years developing different solutions that gives you net new capabilities that just weren't possible.

I'll give you one concrete example. This is a story I love telling because a lot of customers have workloads that consume a lot of storage, whether it's a file server or a database. Typically, in your on-prem datacenter, you don't really need to worry about egress charges. It's, "free." So, network typically is high bandwidth/low latency. You don't have to think about that. Once you bridge that into a public cloud, it's easy to bring data into a public cloud, but it's extremely difficult to not only bring it out, but now there's an additional cost that you have to think about, bandwidth, latency and all that. So, when we build some of the VMware cloud solutions, we wanted to make that sure it was brought close enough to the hyperscalers to bring in all those services.

But one of the capabilities, as simple as just file sharing, as an example, customers can actually bring in their data, and through the integration and the joint engineer work that we've done there, both the egress and the ingress, there are no charges for that because we've done really interesting engineer work at the fundamental level. So, now, you don't have to think about egress as a constraint for the organization, which really allows for our customers to really think about how best to deliver that application, how best to evolve that application, because now this egress concept where you need to think about if you consume public cloud in the traditional manner, you actually have a net new way to consume it. There are a lot of these examples, where you get to rethink your architecture and new set of capabilities that the business can now push on the IT organizations and deliver, that wasn't possible before if you just purely looked at an on-prem model.

David Linthicum:
So, how is VMware looking at the new trends? I think there's going to be huge, explosive growth in data requirements, just because of the training data and output data that we need for AI systems, certainly generative AI. So, customers out there, corporate America is adopting generative AI for lots of good reasons. They're using it as a core base within their business systems, supply chains, things like that. It does a lot of things better. However, it needs a tremendous amount of processing and a tremendous amount of storage. So, how are you guys looking at that at VMware?

William Lam:
I'm glad you brought that up. This definitely came about even before the whole ChatGPT, and just kind of revolutionized the whole industry and getting everybody to think about AI and how that works. We already saw some of that from our customers. With the supply chain challenges, customers have done lots of interesting work already in the AI/ML space, even just general graphics processing. If you think about the inability to get hardware, the time it took from ordering to rack and stack was anywhere from six [months] to maybe even up to a year just only a few years back. They had to look into the public cloud, and how do I start to do some of the training in the public cloud?

Compute was an area that they were looking into, and having the different types of compute instances really helped customers optimize where they were putting the workloads. Because not every public cloud provider has the same set of regions and locations, customers either decide based on their requirements, global proximity. We talked a little bit about data sovereignty and compliance, and we still have those trends even into these training models. So, we have some organizations that are able to take advantage of these public cloud solutions. They can build that training and then bring those models back into on-prem. We have a lot of other customers and rightfully so. There is a general concern of IP, the hallucinations that we're seeing in a lot of these, where it's trained on such a broad data set. We look at how many organizations are starting to think about AI. I wouldn't say it's well figured out.

We're all doing some interesting experiments and there's lots of stuff that's happening extremely fast, but one thing is clear: Lots of data, lots of compute, lots of graphic processing. How do you optimize that, especially in the current macroeconomics? So, we do see a class of customers that are leveraging compute and the public cloud to do this training, bringing it back into an on-prem environment where the data may not be as sensitive. We do have other customers that say, "You know what? I need to be able to build full stack in my on-prem datacenter." With the growing concerns of IP, intellectual property, and just context of where data—organizations are wanting to have AI/ML done in a smart way, and not necessarily like, "Hey, I'm just going to expose a chatbot and call it a day."

To be able to drive that unique business value, you really need to train on your specific domain. What we're hearing and seeing is that these models should be smaller, to a certain degree, so not in the several terabytes or larger where there's a cost implication. So, I think we'll start to see some optimizations for different verticals and domains, where customers are going to want to train on their own data set, and then expose that functionality, whether they do that directly to their customers or they expose that as a new business model to other organizations to consume. So, I do think on the training front, we are going to see more of that land in on-prem.

There are a lot of innovations happening right now in the LLMs. I think even we have just announced some models that we published. So, I think there will be some interesting work that's going to happen in the next 6 to 12 months, to see where things are going with the training side of it. On the inference side, it's interesting because we do believe that a lot of that is still going to be residing—the compute and the function of inferencing will still reside at the edge. So, you're only going to see that continue to expand.

We've had different organizations like marketplaces, food vendors that are wanting to optimize their operations. So, being able to look at how much capacity they have, whether we're talking about food or groceries, a lot of that inferencing, the hardware, the gear, that was all done locally, and now we're just only going to see an increase of that compute happening there. Then how do you bring in all this together? So, VMware I think is well-positioned to help support our customers on this hybrid journey in providing this consistent interface, which is what we call VMware Cloud. I'm sure you'll hear a lot more interesting and exciting announcements.

David Linthicum:
Where can we find you on the Web, work you've done? How would people typically reach out to you?

William Lam:
I'm on a couple of social media channels but I am also on Twitter/X. That's the easiest way to engage with me, on there. I'm also on Mastodon and BlueSky, @lamW. You can find me there. Then I also author my personal blog at WilliamLam.com, where I spend most of my time writing about various technologies, especially around VMware.

David Linthicum:
Yeah, I'm going to run you down once I'm out there. So, I'm looking forward to that a lot, in terms of the talk and multi-cloud, complex cloud architecture, solutions that are out there. I think that's where it's at right now. That's a part of the market that we have to figure out, a huge growth space right now. Certainly with the fuel on the fire with AI and the storage needs for that, It's going to be a wild ride over the next few years. So, if you enjoyed this podcast, like us, rate us, and subscribe. You can also check out our past episodes, including those hosted by my good friend, Mike Kavis. Find out more at DeloitteCloudPodcast.com, all one word. If you'd like to contact me directly, you can e-mail me at dlinthicum@deloitte.com. So, until next time, best of luck with your cloud journey. Stay safe. Cheers.

Operator:

Visit the On Cloud library
www.deloitte.com/us/cloud-podcast